CZASOPISMO TECHNICZNE M
WYDAWNICTWO POLITECHNIKI KRAKOWSKIEJ

z. 3-M/2008
ISSN 0011-4561
ISSN 1897-6328

MAGDALENA KWOLEK*

DEVELOPMENT OF ARTIFICIAL NEURAL NETWORK BASED SPEECH SYNTHESIS FOR THE POLISH LANGUAGE

PROJEKT SYSTEMU KOMPUTEROWEJ SYNTEZY MOWY Z WYKORZYSTANIEM SIECI NEURONOWYCH

Abstract

The paper describes an MLP network that learns to transcribe Polish text to phonemes and defines the process of transcription. The transcription scheme used is SAMPA for the Polish language. The paper also shows mapping of text to binary patterns and the whole process of adaptation patterns for network's requirements. It describes learning process, and learning patterns were provided by professor Krzysztof Marasek from the Polish-Japanese Institute of Information Technology.

Keywords: *speech synthesis*, *Text-to-Speech system, phoneme transcription*, *MLP*, *SAMPA for Polish*

Streszczenie

W niniejszym artykule opisano wykorzystanie sztucznej sieci neuronowej MLP do zamiany tekstu pisanego w języku polskim na fonemy. Zdefiniowano sposób przeprowadzenia transkrypcji fonetycznej. Schemat transkrypcji oparty jest na alfabecie fonetycznym SAMPA dla języka polskiego. Przedstawiono proces przystosowania próbek tekstowych dla potrzeb sieci, czyli zamiany na postać binarną oraz generowanie okna. Opisano również proces uczenia sieci, a jako dane uczące wykorzystano bazę profesora Krzysztofa Maraska z Polsko-Japońskiej Wyższej Szkoły Technik Komputerowych.

Słowa kluczowe: *synteza mowy*, *system TTS*, *transkrypcja fonetyczna*, *MLP*, *SAMPA dla języka polskiego*

*Magdalena Kwolek, V year student, Institute of Applied Informatics, Cracow Unversity of Technology.

1. Introduction

One of the tasks that multimedia systems should handle is a speech synthesis problem. There are a few algorithms in speech synthesis, and all are rule oriented. A disadvantage of such solution is that tens of letter-to-sound (LTS) rules are needed and they have to be compiled by phonetic experts. Moreover, there is a problem of the limited size of words base. Centres of education are making attempts to work out other solutions. Thanks to cooperation of the Institute of Applied Informatics at Cracow University of Technology and the University of Bergen in Norway, an idea was launched to develop artificial neural networks (ANN) for speech synthesis.

The phonetic analysis in a rule based system includes letter-to-sound rules and a dictionary for accurate pronunciation of any word. The dictionary is used as an exceptions list of those words whose pronunciations cannot be predicted on basis of LTS rules alone. In a text-to-speech (TTS) system the list of rules and the dictionary of exceptions are special for each language.

In the present paper an idea of developing an artificial neural network for transcription of Polish text to phonemes has been suggested. The ANN approach to the problem of phoneme transcription can be described as a pattern matching technique where a set of transcription patterns is created to determine how letters are transcribed into phonemes of a given language. Such a solution should perform well on transcription of words not encountered before. This depends, of course, on how well the neural network has been trained. In such an approach what is of essential significance is the learning pattern rather than the size of the words base.

2. Phonetic transcription

A text-to-speech system is composed of two parts. The first one is Natural Language Processing (NLP), which is a process of transformation of letters to their phoneme representation, before their pronunciation. The next step, called Digital Signal Processing (DSP), converts the symbolic linguistic representation into sound.

The main element of NLP process is phonetic transcription which can be described as a visual system of symbolisation of the sounds occurring in the human spoken language.

2.1. SAMPA for Polish

The transcription scheme used in this paper is based on SAMPA (Speech Assessment Methods Phonetic Alphabet). It is compatible with ASCII code, so it makes it easier to type on a typical keyboard. For the Polish language SAMPA contains eight vowels and twenty-nine consonants. It is necessary to add stress mark - in our case it is symbol ‘!’.

The transcription scheme consists of thirty-eight different symbols. Words transcribed using SAMPA notation and its grammatical representations are used as a learning pattern for the artificial neuron network. To ensure the best training result an adequately big base of words is required. Creation of such base is one of the problems for this method. There are a few centres of education in our country which are trying to create such base, but their work is not yet freely available. In this work we used the base created by professor Krzysztof Marasek from the Polish-Japanese Institute of Information Technology.

SAMPA	Orthography	Transcription
i	PIT	pit
I	typ	tIp
e	test	test
a	pat	pat
o	pot	pot
u	puk	puk
e~	gęś	ge~s'
o~	wąs	vo~s
p	pik	pik
b	bit	bit
t	test	test
d	dym	dIm
k	kit	kit
g	gen	gen
f	fan	fan
v	wilk	vilk
s	syk	sIk
z	zbir	zbir
S	szyk	SIk
Z	żyto	ZIto
s'	świt	s'vit
z'	źle	z'le
x	hymn	xImn
ts	cyk	tsIk
dz	dzwon	dzvon
dZ	dżem	dZem
tS	czyn	tSIn
ts'	ćma	ts'ma
dz'	dźwig	dz'vik
m	mysz	mIS
n	nasz	naS
n'	koń	kon'
N	pęk	peNk
l	luk	luk
r	ryk	rIk
w	łyk	wIk
j	jak	jak

Fig. 1. SAMPA symbols for Polish
Rys. 1. Wykaz symboli SAMPA dla języka polskiego

2.2. Structure of transcription

Speech synthesis is difficult to realise on the basis of single letters. Hence it is necessary to produce sequences of letters. These sequences create segments with some coherent characteristics. Literature names such a segment ‘a window’. It is accepted conventionally that window size consists of seven letters [1]. Only the letter in the middle of the window is being predicted and it is active. The other six letters (three on either side of the centre letter) provide the context. To handle the beginning and the end of each word a space is used, indicated by a symbol *. Asterisk (*) is used to fill up the space of seven letters.

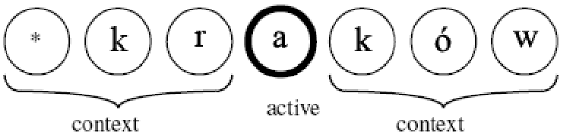

Fig. 2. Word '*kraków' represented in a seven letter window
Rys. 2. Słowo '*kraków' w postaci okna

The words are stepped through the window letter by letter. At each step the network computes the phoneme that corresponds to the letter in the middle of the window. The desired output of the network is the correct phoneme associated with the centre of the window.

3. Transcription model

In Figure 3 the mechanism which is responsible for phonetic transcription is shown. It uses an artificial neural network.

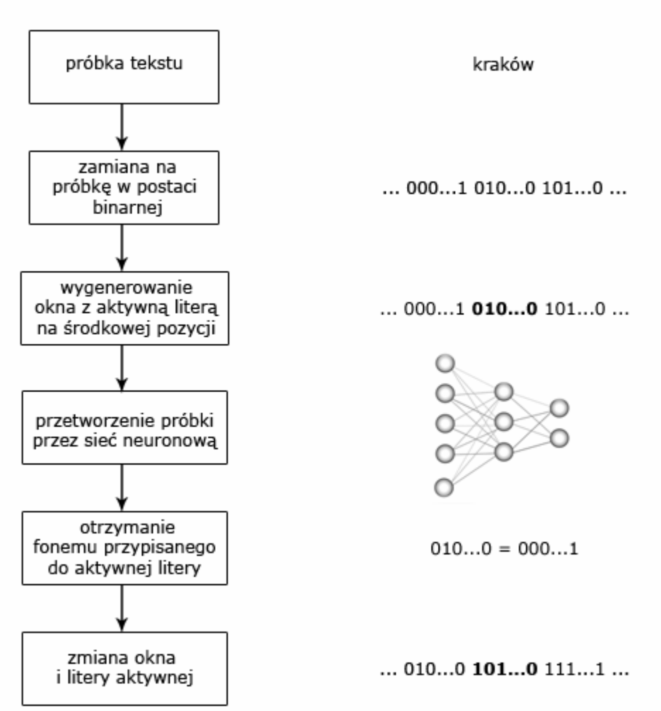
próbka tekstu
kraków
zamiana na próbkę w postaci binarnej
... 000...1 010...0 101...0 ...
wygenerowanie okna z aktywną literą na środkowej pozycji
... 000...1 010...0 101...0 ...
przetworzenie próbki przez sieć neuronową
otrzymanie fonemu przypisanego do aktywnej litery
010...0 = 000...1
zmiana okna i litery aktywnej
... 010...0 101...0 111...1 ...

Fig. 3. Transcription scheme using artificial neural network
Rys. 3. Schemat procesu transkrypcji z wykorzystaniem sieci neuronowych

The network used in this paper is the MLP (Multi Layer Perceptron) which accepts only 0 or 1 as input and output data. Therefore each letter pattern has to be mapped into an equivalent binary pattern that is fed into a neural network. SAMPA for Polish consists of thirty-seven symbols, so the coding table has to consist of this number of items. For instance, character ‘*’ is represented as 37 zeros, and the letter ‘a’ by one (1) in position 37 and zeros elsewhere.

0000000000000000000000000000000000000 – *
000000000000000000000000000000000000**1** – a
00000000000000000000000000000000000**1**0 – ą…

In Figure 4 the pattern ‘*kraków’ is shown in a binary format.

0000000000000000000000000000000000000 – *
00000000000000000000000**1**0000000000000 – k
00000000000000**1**0000000000000000000000 – r
000000000000000000000000000000000000**1** – a
00000000000000000000000**1**0000000000000 – k
0000000000000000**1**00000000000000000000 – ó
0000000001000000000000000000000000000 – w

Fig. 4. The input pattern ‘*kraków’ in binary format
Rys. 4. Wejściowa próbka tekstowa „*kraków” w reprezentacji binarnej

Phonemes also have to be mapped into a binary pattern. It is necessary to add one position to phoneme’s binary notation. On that position stress will be marked.

The next step is to generate a window and to mark the active letter. At each step of processing the pattern ‘*kraków’ window will change as shown in Fig. 5.

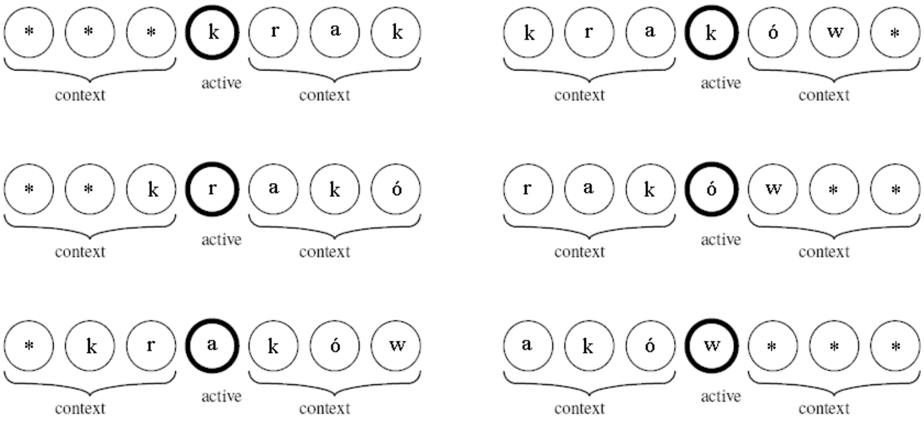

Fig. 5. Changes of a window during processing the pattern ‘*kraków’
Rys. 5. Postać okna podczas analizy próbki ‘*kraków’

Data in such form will be fed into neural network. The MLP network used in this paper consists of 3 layers of processing units. Connections in such network are limited to one direction such that the activations of the input neurons are updated first, followed by hidden layers, and then finished with the outputs.

Because SAMPA for the Polish language consists of thirty-seven symbols and the window size is 7, the input layer is represented by an array of 37 x 7 (259) neurons. The output layer consists of 38 neurons, 37 to represent different phonemes and one more to include stress. To find the correct number of neurons in a hidden layer experiments have to be done.

Each neuron input has an associated weight (w_1, w_2, ..., w_n). Each signal x_j is multiplied by an adequate value that is weight w_{ij}. Weight affects the perception of the given input signal and its participation in creating the output signal by neuron. Added products of signals and weights make up the argument of activation function $f(s_i)$.

The output of each neuron is thus

$$y_i = f\left(\sum_{j-1}^{N} w_{ij} x_j\right) \tag{1}$$

where:

x_i – input signals,
y_i – output signal,
w_{ij} – weights.

The activation function of perceptron is nonlinear and sigmoidal. It means that the output signal can accept only 1or 0 according to formula

$$y_i(s_i) = \begin{cases} 1 \quad \text{dla} \quad s_i > 0 \\ 0 \quad \text{dla} \quad s_i \le 0 \end{cases} \tag{2}$$

where s_i is the output signal of adder

$$s_i = \sum_{j=0}^{N} w_{ij} x_j \tag{3}$$

In this formula it was assumed that vector x of length equal N was widen with zero component $x_0 = 1$ which makes the polarisation signal.

3.1. MLP learning technique

The created mechanism, responsible for phonetic transcription, will be worked out in a neural network learning process.

The learning technique for perceptron is the back – propagation technique. With initial assigned weights w_{ij} learning vector x is fed into the network and output signal is computed. As a result of comparison actual value y_i and correct answer d_i an update of weights is made using formula

$$\Delta w_{ij} = x(d_i - y_i) \tag{4}$$

in accordance with guidelines:

$y_i = d_i$ – weights stay without changes,
$y_i = 0$, $d_i = 1$ – weights update in accordance with formula $w_{ij}(n+1) = w_{ij}(n) + x_j$,
$y_i = 1$, $d_i = 0$ – weights update in accordance with formula $w_{ij}(n+1) = w_{ij}(n) - x_j$,

After update of weights new learning vector and correct answer d_i are fed into the network. Update of weights is made again. This process is repeated until differences between output and correct answer are minimised. A characteristic feature of this method is that during the learning process only information about the actual value of output and correct value are used.

4. Summary

A new approach to speech synthesis problem, described in the paper, allows creation of a system not limited by included words base. This method also allows the use of transcription mechanism for any language or dialect, only the pattern will vary. It has an influence on software developing process, which will not depend on experts in linguistics.

Such a speech synthesis system using artificial neural network was developed for Norwegian. The transcription scheme consisted of fifty-six different units. By choosing the window size of 7 letters the input layer was represented by an array of 392 neurons. As a result of experiments the selected size of hidden layer was 80. The MLP network have been trained by a back propagation algorithm on 32 000 words. The number of iterations was set on 50. To check if a feed forward network is capable of automatically hyphenating Polish, two main experiments were performed. A small-scale experiment showed that an MPL network is able to transcribe Polish five-letter words to their phonetic representation very well. The network ability to generalise was tested on an unknown dictionary of 1000 words. The target phoneme string and the produced phoneme string were identical in 85%. In the second experiment a sample of about 4500 Norwegian words of any length were trained. The performance of the network was tested on a dictionary of 1000 unknown and known words and estimated at 93%. The target phoneme string and the produced phoneme string were aligned and compared character by character. However, only a limited number of Norwegian phonemes (39) was studied in both experiments.

References

[1] Kristensen T., Treeck B., Falck-Olsen R., *Phoneme Transcription of Norwegian Text*, 13th International Conference on Artificial Neural Network and International Conference on Neural Information Processing, ICANN/ICONIP, Istanbul, Turkey 2003.

[2] Tadeusiewicz R., *Sieci neuronowe*, wyd. 2, Akademicka Oficyna Wydawnicza RM, Warszawa 1993.

[3] http://www.phon.ucl.ac.uk/home/sampa/polish.htm.